PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

B3

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12N 15/10, C12Q 1/68 // C12N 9/00

A1

(11) International Publication Number: WO 98/41622

(43) International Publication Date: 24 September 1998 (24.09.98)

(21) International Application Number: PCT/DK98/00103

(22) International Filing Date: 18 March 1998 (18.03.98)

(30) Priority Data:

0306/97	18 March 1997 (18.03.97)	DK
0433/97	17 April 1997 (17.04.97)	DK
0623/97	30 May 1997 (30.05.97)	DK

(71) Applicant: NOVO NORDISK A/S [DK/DK]; Novo Allé, DK-2880 Bagsværd (DK).

(72) Inventors: BORCHERT, Torben, Vedel; Novo Nordisk a/s, Novo Allé, DK-2880 Bagsværd (DK). KRETZSCHMAR, Titus; Kasperlmühlstrasse 6, D-81739 Munich (DE). CHERRY, Joel, R.; 916 Anderson Road, Davis, CA 95616 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, GW, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: METHOD FOR CONSTRUCTING A LIBRARY USING DNA SHUFFLING

(57) Abstract

A method for the construction of a library of recombined polynucleotides from a number of different starting single or double stranded parental DNA templates is disclosed, wherein the starting single or double stranded parental DNA templates represent discrete points in a population of genes encoding evolutionary or synthetic homologues of a peptide having homologies ranging over a broad spectrum from less than 15 % to more than 80 %, said population exhibiting at least one identification sequence, and whereby said genes are subjected to a gene shuffling procedure to generate shuffled mutants of said population of genes representing additional discrete points between those of said starting templates.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

Title: METHOD FOR CONSTRUCTING A LIBRARY USING DNA SHUFFLING

FIELD OF THE INVENTION

The present invention relates to optimizing DNA sequences in order to (a) improve the properties of a protein of interest by artificial generation of genetic diversity of genes encoding proteins having a biological activity of interest by the use of the so-called gene- or DNA shuffling technique to create a large library of "genes", expressing said library of genes in a suitable expression system and screening the expressed proteins in respect of specific characteristics to determine such proteins exhibiting desired properties or (b) improve the properties of regulatory elements such as promoters or terminators by generation of a library of these elements, transforming suitable hosts therewith in operable conjunction with a structural gene, expressing said structural gene and screening for desirable properties in the regulatory element.

BACKGROUND OF THE INVENTION

It is generally found that a protein performing a certain bioactivity exhibits a certain variation between genera and even between members of the same species differences may exist. This variation is of course even more outspoken at the genomic level.

This natural genetic diversity among genes coding for proteins having basically the same bioactivity has been generated in Nature over billions of years and reflects a natural optimization of the proteins coded for in respect of the environment of the organism in question.

In today's society the conditions of life are vastly removed from the natural environment and it has been found that the naturally occurring bioactive molecules are not optimized for the various uses to which they are put by mankind, especially when they are used for industrial purposes.

It has therefore been of interest to industry to identify such bioactive proteins that exhibit optimal properties in respect of the use to which it is intended.

This has for many years been done by screening of natural sources, or by use of mutagenesis. For instance, within the technical field of enzymes for use in *e.g.* detergents, the washing and/or dishwashing performance of *e.g.* naturally occurring proteases, lipases, amylases and cellulases have been improved significantly, by *in vitro* modifications of the enzymes.

In most cases these improvements have been obtained by site-directed mutagenesis resulting in substitution, deletion or insertion of specific amino acid residues which have been chosen either on the basis of their type or on the basis of their location in the secondary or tertiary structure of the mature enzyme (see for instance US patent no. 4,518,584).

In this manner the preparation of novel polypeptide variants and mutants, such as novel modified enzymes with altered characteristics, *e.g.* specific activity, substrate specificity, thermal, pH and salt stability, pH-optimum, pI, K_m, V_{max} etc., has successfully been performed to obtain polypeptides with improved properties.

For instance, within the technical field of enzymes the washing and/or dishwashing performance of *e.g.* proteases, lipases, amylases and cellulases have been improved significantly.

An alternative general approach for modifying proteins and enzymes has been based on random mutagenesis, for instance, as disclosed in US 4,894,331 and WO 93/01285

As it is a cumbersome and time consuming process to obtain polypeptide variants or mutants with improved functional properties a few alternative methods for rapid preparation of modified polypeptides have been suggested.

Weber et al., (1983), Nucleic Acids Research, vol. 11, 5661-5661, describes a method for modifying genes by *in vivo* recombination between two homologous genes. A linear DNA sequence comprising a plasmid vector flanked by a DNA sequence encoding alpha-1 human interferon in the 5'-end and a DNA sequence encoding alpha-2 human interferon in the 3'-end is constructed and transfected into a rec A positive strain of *E. coli*. Recombinants were identified and isolated using a resistance marker.

Pompon *et al.*, (1989), Gene 83, p. 15-24, describes a method for shuffling gene domains of mammalian cytochrome P-450 by *in vivo* recombination of partially homologous sequences in *Saccharomyces cerevisiae* by transforming *Saccharomyces cerevisiae* with a linearized plasmid with filled-in ends, and a DNA fragment being partially homologous to the ends of said plasmid.

In WO 97/07205 a method is described whereby polypeptide variants are prepared by shuffling different nucleotide sequences of homologous DNA sequences by *in vivo* recombination using plasmid DNA as template.

US patent no. 5,093,257 (Assignee: Genencor Int. Inc.) discloses a method for producing hybrid polypeptides by *in vivo* recombination. Hybrid DNA sequences are produced by forming a circular vector comprising a replication sequence, a first DNA sequence encoding the amino-terminal portion of the hybrid po-

lypeptide, a second DNA sequence encoding the carboxy-terminal portion of said hybrid polypeptide. The circular vector is transformed into a rec positive microorganism in which the circular vector is amplified. This results in recombination of said circular vector mediated by the naturally occurring recombination mechanism of the rec positive microorganism, which include prokaryotes such as *Bacillus* and *E. coli*, and eukaryotes such as *Saccharomyces cerevisiae*.

One method for the shuffling of homologous DNA sequences has been described by Stemmer (Stemmer, (1994), Proc. Natl. Acad. Sci. USA, Vol. 91, 10747-10751; Stemmer, (1994), Nature, vol. 370, 389- 391). The method concerns shuffling homologous DNA sequences by using *in vitro* PCR techniques. Positive recombinant genes containing shuffled DNA sequences are selected from a DNA library based on the improved function of the expressed proteins.

The above method is also described in WO 95/22625. WO 95/22625 relates to a method for shuffling of homologous DNA sequences. An important step in the method described in WO 95/22625 is to cleave the homologous template double-stranded polynucleotide into random fragments of a desired size followed by homologously reassembling of the fragments into full-length genes.

A disadvantage inherent to the method of WO 95/22625 is, however, that the diversity generated through that method is limited due to the use of homologous gene sequences (as defined in WO 95/22625).

Another disadvantage in the method of WO 95/22625 lies in the production of the random fragments by the cleavage of the template double-stranded polynucleotide.

A further reference of interest is WO 95/17413 describing a method of gene or DNA shuffling by recombination of specific

DNA sequences - so-called design elements (DE) - either by recombination of synthesized double-stranded fragments or recombination of PCR generated sequences to produce so-called functional elements (FE) comprising at least two of the design elements. According to the method described in WO 95/17413 the recombination has to be performed among design elements that have DNA sequences with sufficient sequence homology to enable hybridization of the different sequences to be recombined.

WO 95/17413 therefore also entails the disadvantage that the diversity generated is relatively limited. Furthermore the methods described are time consuming, expensive, and not suited for automation.

Despite the existence of the above methods there is still a need for better iterative *in vivo* recombination methods for preparing novel polypeptide variants. Such methods should also be capable of being performed in small volumes, and amenable to automation.

Furthermore, there also is a need for methods providing the possibility of being able to shuffle genes with relatively low homology.

SUMMARY OF THE INVENTION

The present invention relates to a method for the construction of a library of recombined polynucleotides from a number of different starting single or double stranded parental DNA templates, wherein said starting single or double stranded parental DNA templates represent discrete points in a population of genes encoding evolutionary or synthetic homologues of a peptide having homologies ranging over a broad spectrum from less than 15% to more than 80%, said population exhibiting at least one identification sequence, and whereby said genes are subjected to a gene shuffling procedure to generate shuffled

mutants of said population of genes representing additional discrete points between those of said starting templates.

The gene shuffling procedure to be used according to the invention can be any suitable method such as those described above or a procedure as described in our co-pending patent application filed on the same date, and outlined below.

According to that procedure template shifts of newly synthesized DNA strands during *in vitro* DNA synthesis are utilized to achieve DNA shuffling.

In a further aspect the invention relates to a method of identifying polypeptides exhibiting improved properties in comparison to naturally occurring polypeptides of the same bioactivity, whereby a library of recombined polynucleotides produced by the above process are cloned into an appropriate vector, said vector is then transformed into a suitable host system, to be expressed into the corresponding polypeptides, said polypeptides are then screened in a suitable assay, and positive results selected.

In a still further aspect the invention relates to a method for producing a polypeptide of interest as identified in the preceding process, whereby a vector comprising a polynucleotide encoding said polypeptide is transformed into a suitable host, said host is grown to express said polypeptide, and the polypeptide recovered and purified.

DEFINITIONS

Prior to discussing this invention in further detail, the following terms will first be defined.

"Shuffling": The term "shuffling" herein means recombination of nucleotide sequence fragment(s) between two or more polynucleotides resulting in output polynucleotides (i.e.

polynucleotides having been subjected to a shuffling cycle) having a number of nucleotide fragments exchanged, in comparison to the input polynucleotides (i.e. starting point polynucleotides).

"Homology of DNA sequences or polynucleotides" In the present context the degree of DNA sequence homology is determined as the degree of identity between two sequences indicating a derivation of the first sequence from the second. The homology may suitably be determined by means of computer programs known in the art, such as GAP provided in the GCG program package (Program Manual for the Wisconsin Package, Version 8, August 1994, Genetics Computer Group, 575 Science Drive, Madison, Wisconsin, USA 53711)(Needleman, S.B. and Wunsch, C.D., (1970), Journal of Molecular Biology, 48, 443-453).

"Homologous": The term "homologous" means that one single-stranded nucleic acid sequence may hybridize to a complementary single-stranded nucleic acid sequence. The degree of hybridization may depend on a number of factors including the amount of identity between the sequences and the hybridization conditions such as temperature and salt concentration as discussed later (*vide infra*).

Using the computer program GAP (*vide supra*) with the following settings for DNA sequence comparison: GAP creation penalty of 5.0 and GAP extension penalty of 0.3, it is in the present context believed that two DNA sequences will be able to hybridize (using low stringency hybridization conditions as defined below) if they mutually exhibit a degree of identity preferably of at least 70%, more preferably at least 80%, and even more preferably at least 85%.

"heterologous": If two or more DNA sequences mutually exhibit a degree of identity which is less than above specified, they are in the present context said to be "heterologous".

"Hybridization:" Suitable experimental conditions for determining if two or more DNA sequences of interest do hybridize or not is herein defined as hybridization at low stringency as described in detail below.

Molecules to which the oligonucleotide probe hybridizes under these conditions are detected using a x-ray film or a phosphoimager.

"primer": The term "primer" used herein especially in connection with a PCR reaction is an oligonucleotide (especially a "PCR-primer") defined and constructed according to general standard specification known in the art ("PCR A practical approach" IRL Press, (1991)).

"A primer directed to a sequence:" The term "a primer directed to a sequence" means that the primer (preferably to be used in a PCR reaction) is constructed to exhibit at least 80% degree of sequence identity to the sequence part of interest, more preferably at least 90% degree of sequence identity to the sequence part of interest, which said primer consequently is "directed to". The primer is designed in order to specifically anneal at the region at a given temperature it is directed towards. Especially identity at the 3' end of the primer is essential for the function of the polymerase, i.e. the ability of a polymerase to extend the annealed primer.

"Flanking" The term "flanking" used herein in connection with DNA sequences comprised in a PCR-fragment means the outermost partial sequences of the PCR-fragment, both in the 5` and 3` ends of the PCR fragment.

"Polypeptide" Polymers of amino acids sometimes referred to as protein. The sequence of amino acids determines the folded conformation that the polypeptide assumes, and this in turn determines biological properties such as activity. Some polypeptides consist of a single polypeptide chain

(monomeric), whilst other comprise several associated polypeptides (multimeric). All enzymes and antibodies are polypeptides.

"Enzyme" A protein capable of catalysing chemical reactions. Specific types of enzymes are a) hydrolases including amylases, cellulases and other carbohydrases, proteases, and lipases, b) oxidoreductases, c) Ligases, d) Lyases, e) Isomerases, f) Transferases, etc. Of specific interest in relation to the present invention are enzymes used in detergents, such as proteases, lipases, cellulases, amylases, etc.

DETAILED DESCRIPTION OF THE INVENTION

All possible genes encoding a polypeptide of the same evolutionary origin can be seen as a very large population of DNA sequences (*e.g.* {G_{sp} | the set of genes encoding a serine protease}). It has been found that the homology between the polypeptides encoded by single members of such a population may be even as small as less than 15% (the genes originating from "distant" organisms).

When searching for polypeptides suited for the various purposes that mankind has developed, it has been found difficult, if not impossible at our present level of knowledge, to conclude in a rational manner on the optimal configuration of the polypeptide in question. Therefore it was found desirable to provide a simple method of generating a sub-population of the above mentioned very large population, but representing a substantial part of the variation possible within the large population.

The object of the present invention is thus to provide a method whereby it is possible to shuffle components of genes

encoding polypeptides of the same functionality, but having only low homologies.

To this end it is necessary to obtain a reasonable knowledge of the population in question, meaning having at disposal a number of individual members (*e.g.* 5, 10, 15 or more members) representing as high a variation as possible. This small sub-population is then used as a starting point for generating a much larger sub-population of genes. The corresponding polypeptides of the large sub-population obtained are then displayed and screened in an appropriate manner to identify such members of the large sub-population that are optimal for the intended purpose.

It was found that the expansion of the starting sub-population to the large sub-population could be accomplished using gene shuffling methods.

Such methods as described in the literature provide means to exchange DNA fragments between genes coding for polypeptides of a reasonably high homology, typically to be above 80%, resulting in the generation of novel genes encoding polypeptides having homologies between 80% and 99%.

It was also found that in the method of the invention it was necessary as starting population to use genes encoding polypeptides that are at least from 70% to 80% homologous to at least one other gene in the starting population.

According to the invention it is thus important to start from a population or sub-set of genes which comprises intermediate sequences ranging from genes being rather similar to genes being rather dissimilar, but still having the same evolutionary origin (function). Only then a shuffling of even rather heterologous sequences is feasible. The stepwise shuffling of, first, quite homologous genes creates new species which are not contained in the starting population, and which, in the subse-

quent shuffling rounds, will recombine with each other and with other more heterologous genes from the starting population, and so on.

Finally, hybrids are generated in which sequence parts from very heterologous starting genes can be found. These retrieved starting genes would have never been shuffled without having the intermediate species in the starting population because of a too large "sequence space" distance.

Having this condition fulfilled it was found that it was possible to generate genes encoding novel functional polypeptides having a homology as low as the minimum degree of homology represented in the starting population. In principle the homology range in the final population may be even greater than that for the starting population.

The present invention relates in its first aspect to a method for the construction of a library of recombined polynucleotides from a number of different starting single or double stranded parental DNA templates, wherein said starting single or double stranded parental DNA templates represent discrete points in a population of genes encoding evolutionary or synthetic homologues of a peptide having homologies ranging over a broad spectrum from less than 15% to more than 80%, said population exhibiting at least one identification sequence, and whereby said genes are subjected to a gene shuffling procedure to generate shuffled mutants of said population of genes representing additional discrete points between those of said starting templates.

According to the invention it is possible to use parental DNA templates representing homologies ranging from less than 45%, 40%, 35%, 30%, 25%, 20%, or 15% to more than 80%, 85%, 90%, 95%, or 99%.

In specific embodiments at least one identification sequence is identified and primers constructed to anneal thereto. These sequences can be located anywhere on the genes.

In a preferred embodiment at leasts two identification sequences are identified. These sequences can be located at any distance from each other, but it is preferred that they are located as far as possible from each other on the genes.

According to these embodiments said identification sequences may correspond to an amino acid sequence of from 4 to 8 amino acid residues, which sequence is highly conserved among the peptides encoded by the collection of starting single or double stranded parental DNA templates, preferably from 5 to 7 amino acid residues.

It is preferred that the identification sequences are located a distance apart corresponding to the average size of the genes in said collection with a variation of up to 40%. The longer apart the sequences are the larger a part of the gene is shuffled.

However, situations may arise, where it is desired only to shuffle the sequences between identification sequences located quite close to each other.

As indicated above the gene shuffling method used in the method of the invention is of less or no significance. In principle any method will work.

Thus the methods disclosed in WO 95/22625 and WO 95/17413 are fully operable in the present invention. Details showing how these methods may be used for practising the present invention are indicated in the Examples below.

Therefore further gene shuffling methods described in co-filed patent applications are also contemplated for use in the present method.

According to one of these procedures template shifts of newly synthesized DNA strands during *in vitro* DNA synthesis are utilized to achieve DNA shuffling.

More specifically that method provides for the construction of a library of recombined homologous polynucleotides from a number of different starting single or double stranded parental DNA templates and primers by induced template shifts during an *in vitro* polynucleotide synthesis using a polymerase, whereby

A. extended primers or polynucleotides are synthesized by
   a) denaturing parental double stranded DNA templates to produce single stranded templates,
   b) annealing said primers to the single stranded DNA templates,
   c) extending said primers by initiating synthesis by use of said polymerase,
   d) cause arrest of the synthesis, and
   e) denaturing the double strand to separate the extended primers from the templates,

B. a template shift is induced by
   a) isolating the newly synthesized single stranded extended primers from the templates and repeating steps A.b) to A.e) using said extended primers produced in (A) as both primers and templates, or
   b) repeating steps A.b) to A.e),

C. the above process is terminated after an appropriate number of cycles of process steps A. and B.a), A. and B.b), or combinations thereof, and

D. optionally the produced polynucleotides are amplified in a standard PCR reaction with specific primers to selectively amplify polynucleotides of interest.

In a further specific embodiment the gene shuffling is performed by the method described in our co-filed application, whereby conserved regions of heterologous DNA sequences are identified for shuffling of heterologous DNA sequences of interest having at least one conserved region comprising the following steps:

i) One or more conserved region(s) (designated "A,B,C" etc..) in two or more of the heterologous sequences are identified.

ii) Two sets of PCR primers (each set comprising a sense and an anti-sense primer) for one or more conserved region(s) identified in (i) are constructed.

In these primers, one set (named: "a"=sense primer; "a'"=anti-sense primer) is directed to a sequence region 5' (sense strain) of the conserved region (e.g. conserved region "A"), and the second set (named: "b"=sense primer; "b'"=anti-sense primer) is directed to a sequence region 3' (sense strain) of the conserved region (e.g. conserved region "A"), and the antisense primer "a'" and the sense primer "b" have a homologous sequence overlap of at least 10 base pairs (bp) within the conserved region.

iii) for one or more identified conserved region of interest in step (i) two PCR amplification reactions are performed using the heterologous DNA sequences from step (i) as templates, whereby one of the PCR reactions is using the 5' primer set identified in step (ii) (e.g. named "a","a'") and the second PCR reaction is using the 3' primer set identified in step (ii) (e.g. named "b","b'").

iv) The PCR fragments generated as described in step (iii) for one or more of the identified conserved region in step (i); are isolated.

v) Two or more PCR fragments isolated from step (iv) and performance of a Sequence overlap extension PCR reaction (SOE-PCR) using the isolated PCR fragments as templates are pooled.

vi) The PCR fragment(s) obtained in step (v) are isolated, whereby the isolated PCR fragment comprise numerous different shuffled sequences containing a shuffled mixture of the PCR fragments isolated in step (iv).

In specific embodiments various modifications can be made in the process of the invention. For example it is advantageous to apply a defective polymerase either an error-prone polymerase to introduce mutations in comparison to the templates, or a polymerase that will discontinue the polynucleotide synthesis prematurely to effect the arrest of the reaction.

According to a specific embodiment the peptide is a protease, especially a subtilase.

In the case of a subtilase identification sequences may be located around the aspartic acid in position 32, or the histidine in position 64 and the active serine in position 221 of subtilisin BPN'.

In a further embodiment the peptide is an amylase, especially an α-amylase.

In that case of identification sequences may be located around the Asp in position 100 and the Asp in position 328 of *B.licheniformis* α-amylase.

For α-amylases from Bacillus species the identification sequences may preferentially be located around Tyr in position 8 and around Ser in position 476.

In further embodiments the peptide is a lipase, or a cellulase.

In respect of lipases, suitable identification sequences may be found by using the lipase alignment shown in A. Svendsen *et al.* (1995): Biochemical properties of cloned lipases from the Pseudomonas family, *Biochimica et Biophysica Acta* **1259** 9-17. Examples could be around the Pro in position 10 or around the His in position 285 (using *P. glumae* lipase numbering).

In respect of cellulases, in particular cellulases from family 45 cellulases (see WO 96/29397), suitable identification sequences may be the conserved region ""Thr Arg Tyr Trp Asp Cys Cys Lys Pro/Thr" and the conserved region ""Trp Arg Phe/Tyr Asp Trp Phe". For further details relating to those cellulase identification sequences reference is made to (PCT DK97/00216). See in particular in example 3 of (PCT DK97/00216).

In respect of xylanases, in particular xylanases from family 11 xylanases, suitable identification sequences may be the conserved regions "DGGTYDIY" and "EGYQSSG". For further details relating to those xylanase identification sequences reference is made to (PCT DK97/00216). See in particular in example 1,2 of (PCT DK97/00216).

PCR-primers:

The PCR primers are constructed according to the standard descriptions in the art. Normally they are 10-75 base-pairs (bp) long. However, for the specific embodiment using random or semi-random primers the length may be substantially longer as indicated above.

PCR-reactions:

If not otherwise mentioned the PCR-reaction performed according to the invention are performed according to standard protocols known in the art.

The term "Isolation of PCR fragment" is intended to cover as broad as simply an aliquot containing the PCR fragment. However preferably the PCR fragment is isolated to an extend which remove surplus of primers, nucleotides templates etc..

In an embodiment of the invention the DNA fragment(s) is(are) prepared under conditions resulting in a low, medium or high random mutagenesis frequency.

To obtain low mutagenesis frequency the DNA sequence(s) (comprising the DNA fragment(s)) may be prepared by a standard PCR amplification method (US 4,683,202 or Saiki et al., (1988), Science 239, 487 - 491).

A medium or high mutagenesis frequency may be obtained by performing the PCR amplification under conditions which increase the misincorporation of nucleotides, for instance as described by Deshler, (1992), GATA 9(4), 103-106; Leung et al., (1989), Technique, Vol. 1, No. 1, 11-15.

It is also contemplated according to the invention to combine the PCR amplification (*i.e.* according to this embodiment also DNA fragment mutation) with a mutagenesis step using a suitable physical or chemical mutagenizing agent, *e.g.*, one which induces transitions, transversions, inversions, scrambling, deletions, and/or insertions.

Expressing the recombinant protein from the recombinant shuffled sequences

Expression of the recombinant protein encoded by the shuffled sequence in step vi) of the second and third aspect of the present invention may be performed by use of standard expression vectors and corresponding expression systems known in the art.

Screening and selection

In the context of the present invention the term "positive polypeptide variants" means resulting polypeptide variants possessing functional properties which has been improved in comparison to the polypeptides producible from the corresponding input DNA sequences. Examples, of such improved properties can be as different as *e.g.* enhance or lowered biological activity, increased wash performance, thermostability, oxidation stability, substrate specificity, antibiotic resistance etc.

Consequently, the screening method to be used for identifying positive variants depend on which property of the polypeptide in question it is desired to change, and in what direction the change is desired.

A number of suitable screening or selection systems to screen or select for a desired biological activity are described in the art. Examples are:

Strauberg et al. (Biotechnology 13: 669-673 (1995), describes a screening system for subtilisin variants having a Calcium-independent stability;

Bryan et al. (Proteins 1:326-334 (1986)) describes a screening assay for protease having enhanced thermal stability; and

PCT-DK96/00322 describes a screening assay for lipases having an improved wash performance in washing detergents.

An embodiment of the invention comprise screening or selection of recombinant protein(s), wherein the desired biological activity is performance in dish-wash or laundry detergents. Examples of suitable dish-wash or laundry detergents are disclosed in PCT-DK96/00322 and WO 95/30011.

If the improved functional property of the polypeptide is not sufficiently good after one cycle of shuffling, the polypeptide may be subjected to another cycle.

In an embodiment of the invention wherein polynucleotides representing a number of mutations of the same gene is used as templates at least one shuffling cycle is a backcrossing cycle with the initially used DNA fragment, which may be the wild-type DNA fragment. This eliminates non-essential mutations. Non-essential mutations may also be eliminated by using wild-type DNA fragments as the initially used input DNA material.

Also contemplated to be within the invention is polypeptides having biological activity such as insulin, ACTH, glucagon, somatostatin, somatotropin, thymosin, parathyroid hormone, pituary hormones, somatomedin, erythropoietin, luteinizing hormone, chorionic gonadotropin, hypothalamic releasing factors, antidiuretic hormones, thyroid stimulating hormone, relaxin, interferon, thrombopoeitin (TPO) and prolactin.

It is also contemplated according to the invention to shuffle parental polynucleotides as indicated above originating from wild type organisms of different genera.

The starting parental DNA sequences may be any DNA sequences including wild-type DNA sequences, DNA sequences encoding variants or mutants, or modifications thereof, such as extended or elongated DNA sequences, and may also be the outcome of DNA sequences having been subjected to one or more cycles of shuffling (*i.e.* output DNA sequences) according to the method of the invention or any other method (*e.g.* any of the methods described in the prior art section), or synthetic sequences or otherwise mutagenized sequences.

When using the method of the invention the resulting recombined polynucleotides (*i.e.* shuffled DNA sequences), have had a number of nucleotide fragments exchanged. This results in

replacement of at least one amino acid within the polypeptide variant, if comparing it with the parent polypeptide. It is to be understood that also silent exchanges are contemplated (*i.e.* nucleotide exchange which does not result in changes in the amino acid sequence).

MATERIALS AND METHODS

EXAMPLES

EXAMPLE 1

Shuffling of a pool/population of evolutionary homologues originating from bacterial hosts.

In this Example a gene shuffling method similar to the one described in WO 95/22625 is used:

A population of subtilase-encoding genes or parts of such genes are generated through isolation or by synthesis. Sources for the genes may be as described in Siezen *et al. Protein Engineering* **4** 1991 719-737. The population may also comprise genes encoding the pre-pro subtilases as defined in GenBank entries A13050_1, D26542, A22550, Swiss-Prot entry SUBT_BACAM P00782, and PD498 (Patent Application No. WO 96/34963) with homologies(similarities) ranging from 32% to 64% as calculated by the MegAlign software from DNASTAR Inc. (WI 53715, USA) using the Clustal Method.

The substrates used in the shuffling reaction are represented by linear double stranded DNA generated by PCR amplification using primers located at/directed towards the ends of the DNA to be shuffled. In this instance the primers can conveniently be constructed using the sequences surrounding the

histidine in pos 64 of subtilisin BPN' and the serine in position 221 of subtilisin BPN'. The template for this PCR can either be plasmids containing cloned protease genes or chromosomal DNA extracted from bacterial strains *e.g.* protease secreting bacteria isolated from soil. The substrate will typically be generated separately for all the templates and pooled before the shuffling reaction.

The substrates are fragmented *e.g.* by DNAse I treatment or shearing by sonication as described in WO 95/22625. The generated fragments are separated according to size by agarose gel electrophoresis and generated fragment of the desired size, *e.g.* from 10 to 50 bp. or from 30 to 100 bp, or from 50 to 150 bp, or from 100 to 200 bp are purified from the gel.

These fragments are reassembled by PCR as described in WO95/22625. Optionally, correctly assembled DNA fragments are amplified by subjecting the product from the assembly reaction to another PCR including two primers able to anneal to the ends of correctly assembled fragments. The resulting fragments can be cloned into suitable expression plasmids, and subsequently screened for a specific property, such as thermostability using assays well known in the art.

PATENT CLAIMS

1. A method for the construction of a library of recombined polynucleotides from a number of different starting single or double stranded parental DNA templates, wherein said starting single or double stranded parental DNA templates represent discrete points in a population of genes encoding evolutionary or synthetic homologues of a peptide having homologies ranging over a broad spectrum from less than 15% to more than 80%, said population exhibiting at least one identification sequence, and whereby said genes are subjected to a gene shuffling procedure to generate shuffled mutants of said population of genes representing additional discrete points between those of said starting templates.

2. The method of claim 1, wherein said homologies range from less than 45%, 40%, 35%, 30%, 25%, 20%, or 15% to more than 80%, 85%, 90%, 95%, or 99%.

3. The method of claim 1 or 2, wherein said starting population exhibits at least two identification sequences.

4. The method of any of the claims 1 to 3, wherein said identification sequences corresponds to amino acid sequences of from 4 to 8 amino acid residues, which sequence is highly conserved among the peptides encoded by said collection of starting single or double stranded parental DNA templates, preferably from 5 to 7 amino acid residues.

5. The method of claim 3 or 4, wherein said identification sequences are located a distance apart corresponding to the av-

erage size of the genes in said collection with a variation of up to 40%.

6. The method of claim 3, wherein said variation is 20%, 15%, 10%, or 5%.

7. A method of identifying a polypeptide of interest exhibiting improved properties in comparison to naturally occurring or other known polypeptides of the same activity, whereby a population of recombined polynucleotides produced by a process according to any of the claims 1 to 6 are cloned into an appropriate vector, said vector is transformed into a suitable host system, to be expressed into the corresponding polypeptides, and said polypeptides are screened in a suitable assay, and positive polypeptides selected.

8. A method for producing a polypeptide of interest as identified according to claim 7, whereby a vector comprising a polynucleotide encoding said identified polypeptide is transformed into a suitable host, said host is grown to express said polypeptide, and the polypeptide recovered and purified.

9. The method of claim 8, wherein said peptide is a protease, especially a subtilase.

10. The method of claim 9, wherein said identification sequences are located around the histidine in position 64 and the active serine in position 221 of subtilisin BPN'.

11. The method of claim 8, wherein said peptide is an amylase, especially an α-amylase.

12. The method of claim 11, wherein said identification sequences are located around the Asp in position 100 and the Asp in position 328 of *B.licheniformis* α-amylase.

12. The method of claim 11, wherein said identification sequences are located around the Tyr in position 8 and around Ser in position 476 of *B.licheniformis* α-amylase.

13. The method of claim 8, wherein said peptide is a lipase.

14. The method of claim 13, wherein said identification sequences are located around the Pro in position 10 and around the His in position 285 of *P. glumae* lipase.

15. The method of claim 8, wherein said peptide is a cellulase.

15. The method of claim 8, wherein said peptide is a xylanase.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/DK 98/00103

A. CLASSIFICATION OF SUBJECT MATTER

IPC6: C12N 15/10, C12Q 1/68 // C12N 9/00
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC6: C12N, C07K, C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

SE,DK,FI,NO classes as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 9522625 A1 (AFFYMAX TECHNOLOGIES N.V.), 24 August 1995 (24.08.95), page 9, line 1 - line 16; page 16, line 24 - line 29; page 78, line 16 - line 25, claims --	1-15
P,X	Nature, Volume 391, January 1998, Andreas Crameri et al, "DNA shuffling of a family of genes from diverse species accelerates directed evolution" page 288 - page 291 --	1-15
P,X	WO 9735966 A1 (MAXYGEN, INC.), 2 October 1997 (02.10.97), figure 1 -- --------	1-15

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" erlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
30 June 1998	03 -07- 1998
Name and mailing address of the ISA/ Swedish Patent Office Box 5055, S-102 42 STOCKHOLM Facsimile No. +46 8 666 02 86	Authorized officer Patrick Andersson Telephone No. +46 8 782 25 00

INTERNATIONAL SEARCH REPORT

Information on patent family members

09/06/98

International application No.
PCT/DK 98/00103

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9522625 A1	24/08/95	AU 2971495 A	04/09/95
		CA 2182393 A	24/08/95
		CN 1145641 A	19/03/97
		EP 0752008 A	08/01/97
		JP 10500561 T	20/01/98
		US 5605793 A	25/02/97
WO 9735966 A1	02/10/97	AU 2337797 A	17/10/97
		AU 2542697 A	17/10/97
		WO 9735957 A	02/10/97
		AU 1087397 A	19/06/97
		WO 9720078 A	05/06/97